

Problem 1.

(ii)

$$L_{\pi_{\theta_1}}(\pi_{\theta}) = \eta(\pi_{\theta_1}) + \sum_s d_{\mu}^{\pi_{\theta_1}}(s) \cdot \sum_a \pi_{\theta}(a|s) \cdot A^{\pi_{\theta_1}}(s, a)$$

$$\Rightarrow \nabla_{\theta} L_{\pi_{\theta_1}}(\pi_{\theta}) = \sum_s d_{\mu}^{\pi_{\theta_1}}(s) \sum_a \nabla_{\theta} \pi_{\theta}(a|s) \cdot A^{\pi_{\theta_1}}(s, a)$$

$$\text{又, } \eta(\pi_{\theta}) = \eta(\pi_{\theta_1}) + \sum_s d_{\mu}^{\pi_{\theta_1}}(s) \cdot \sum_a \pi_{\theta}(a|s) \cdot A^{\pi_{\theta_1}}(s, a)$$

$$\Rightarrow \nabla_{\theta} \eta(\pi_{\theta}) = \sum_s \nabla_{\theta} \left(d_{\mu}^{\pi_{\theta_1}}(s) \cdot \sum_a \pi_{\theta}(a|s) \cdot A^{\pi_{\theta_1}}(s, a) \right)$$

$$= \sum_s \left[\left(\nabla_{\theta} d_{\mu}^{\pi_{\theta_1}}(s) \right) \cdot \left(\sum_a \pi_{\theta}(a|s) \cdot A^{\pi_{\theta_1}}(s, a) \right) + d_{\mu}^{\pi_{\theta_1}}(s) \cdot \left(\sum_a \nabla_{\theta} \pi_{\theta}(a|s) \cdot A^{\pi_{\theta_1}}(s, a) \right) \right]$$

$$\Rightarrow \nabla_{\theta} \eta(\pi_{\theta}) \Big|_{\theta=\theta_1} = \sum_s \left[\left(\nabla_{\theta} d_{\mu}^{\pi_{\theta_1}}(s) \right) \cdot \left(\sum_a \pi_{\theta_1}(a|s) \cdot A^{\pi_{\theta_1}}(s, a) \right) + d_{\mu}^{\pi_{\theta_1}}(s) \cdot \left(\sum_a \nabla_{\theta} \pi_{\theta}(a|s) \cdot A^{\pi_{\theta_1}}(s, a) \right) \right]$$

$$= \sum_s d_{\mu}^{\pi_{\theta_1}}(s) \cdot \sum_a \nabla_{\theta} \pi_{\theta}(a|s) \cdot A^{\pi_{\theta_1}}(s, a)$$

$$= \nabla_{\theta} L_{\pi_{\theta_1}}(\pi_{\theta})$$

(i)

$$L_{\pi_{\theta_1}}(\pi_{\theta_1}) = \eta(\pi_{\theta_1}) + \sum_s d_{\mu}^{\pi_{\theta_1}}(s) \sum_a \pi_{\theta_1}(a|s) A^{\pi_{\theta_1}}(s, a)$$

$$\approx \eta(\pi_{\theta_1}) + (\eta(\pi_{\theta_1}) - \eta(\pi_{\theta_1})) = \eta(\pi_{\theta_1})$$

Problem 2:

$$D(\lambda) := \min_{\theta \in \mathbb{R}^d} \left\{ -(\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k})^T (\theta - \theta_k) + \lambda \left(\frac{1}{2} (\theta - \theta_k)^T H (\theta - \theta_k) - \delta \right) \right\},$$

H is a positive definite matrix and hence $L(\theta, \lambda)$ is strictly convex.

$$\nabla_{\theta} L(\theta, \lambda) = -(\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k}) + \lambda H(\theta - \theta_k)$$

Since $L(\theta, \lambda)$ is strictly convex, then a point θ^*, λ^* the global minimum

if and only if $\nabla_{\theta} L(\theta, \lambda) \Big|_{\theta=\theta^*, \lambda=\lambda^*} = 0$

$$\Rightarrow \nabla_{\theta} L(\theta^*, \lambda^*) = -(\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k}) + \lambda^* H(\theta^* - \theta_k) = 0$$

$$\Leftrightarrow \theta^* - \theta_k = \frac{1}{\lambda^*} H^{-1} (\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k})$$

$$\begin{aligned} \text{Hence, } D(\lambda) = L(\theta^*, \lambda) &= \frac{1}{\lambda^*} (\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k})^T H^{-1} (\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k}) + \\ &+ \lambda^* \left(\frac{1}{2\lambda^{*2}} (\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k})^T H^{-1} \cdot H \cdot H^{-1} (\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k}) - \delta \right) \\ &= \frac{1}{2\lambda^*} \left((\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k})^T H^{-1} (\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k}) \right) - \lambda^* \delta \quad \# \end{aligned}$$

$$\lambda^* = \sqrt{\frac{(\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k})^T H^{-1} (\nabla_{\theta} L_{\theta_k}(\theta) \Big|_{\theta=\theta_k})}{2\delta}} \quad (b) \quad \alpha = \frac{1}{\lambda^*}$$

λ^* is obtained using the Cauchy-Schwarz inequality

Problem 3 :

(a)

Hyperparameters :

Learning rate : 0.005

step_size=100

gamma=0.9

Loss function :

Policy loss = $\log_prob(action) * advantage$

Value loss = $L1_smooth_loss$

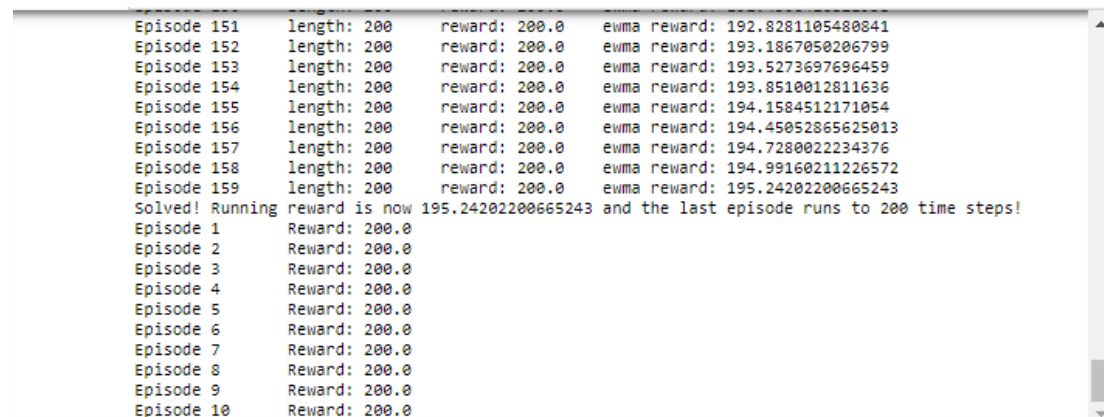
NN architecture :

Policy : Input(4) , FC(128) , relu() , FC(128) , relu() , FC(2) , softmax()

Value : Input(4) , FC(128) , relu() , FC(256) , relu() , FC(1)

Implement result :

Achieve reward threshold at episode 159



```
Episode 151    length: 200    reward: 200.0    ewma reward: 192.8281105480841
Episode 152    length: 200    reward: 200.0    ewma reward: 193.1867050206799
Episode 153    length: 200    reward: 200.0    ewma reward: 193.5273697696459
Episode 154    length: 200    reward: 200.0    ewma reward: 193.8510012811636
Episode 155    length: 200    reward: 200.0    ewma reward: 194.1584512171054
Episode 156    length: 200    reward: 200.0    ewma reward: 194.45052865625013
Episode 157    length: 200    reward: 200.0    ewma reward: 194.7280022234376
Episode 158    length: 200    reward: 200.0    ewma reward: 194.99160211226572
Episode 159    length: 200    reward: 200.0    ewma reward: 195.24202200665243
Solved! Running reward is now 195.24202200665243 and the last episode runs to 200 time steps!
Episode 1      Reward: 200.0
Episode 2      Reward: 200.0
Episode 3      Reward: 200.0
Episode 4      Reward: 200.0
Episode 5      Reward: 200.0
Episode 6      Reward: 200.0
Episode 7      Reward: 200.0
Episode 8      Reward: 200.0
Episode 9      Reward: 200.0
Episode 10     Reward: 200.0
```

(b)

With MC:

Hyperparameters :

Learning rate : 0.01

step_size=100

gamma=0.9

Loss function :

Policy loss = $\log_prob(action) * advantage$

Value loss = $L1_smooth_loss$

NN architecture :

Actor : Input(8) , FC(128) , relu() , FC(2) , softmax()

Critic : Input(8) , FC(128) , relu() , FC(1)

Implement result :

Achieve reward threshold at episode 687

```

Episode 680 length: 337 reward: 200.08770475554302 ewma reward: 195.75910000000000
Episode 681 length: 356 reward: 298.9200389758118 ewma reward: 195.44235399707782
Episode 682 length: 187 reward: 20.279385114327056 ewma reward: 186.68420555294026
Episode 683 length: 231 reward: 277.6023414769363 ewma reward: 191.23011234914003
Episode 684 length: 252 reward: 286.1238837054355 ewma reward: 195.9748009169548
Episode 685 length: 451 reward: 193.73963395912853 ewma reward: 195.86304256906345
Episode 686 length: 517 reward: 187.71417150800391 ewma reward: 195.45559901601047
Episode 687 length: 254 reward: 249.78073029973703 ewma reward: 198.1718555801968

Episode 688 length: 278 reward: 200.08265763324533 ewma reward: 202.2673956828492
Solved! Running reward is now 202.2673956828492 and the last episode runs to 278 time steps!
Episode 1 Reward: 5.833587131155497
Episode 2 Reward: 38.4833789562189
Episode 3 Reward: 242.60193959801694
Episode 4 Reward: 296.3570258473717
Episode 5 Reward: 225.51460402177545
Episode 6 Reward: 251.5885789843295
Episode 7 Reward: 26.860610665735962
Episode 8 Reward: 289.13961619997133
Episode 9 Reward: 236.35794236169806
Episode 10 Reward: 279.1140289801829

```

With TD(0) bootstrapping:

I tried to tune the hyperparameters many times and tried to split actor and critic. But always didn't get a acceptable performance.